Proceedings Paper:
Aloufi, OF, Djemame, K orcid.org/0000-0001-5811-5263, Saeed, F et al. (1 more author) (2021) A survey on predicting workloads and optimizing QoS in the cloud computing. In: Proceedings of the 2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021. 2021 International Congress of Advanced Technology and Engineering, 04-05 Jul 2021, Taiz, Yemen. IEEE , pp. 1-7. ISBN 978-1-6654-2966-5

https://doi.org/10.1109/ICOTEN52080.2021.9493436

A survey on predicting workloads and optimising QoS in the cloud computing

Omar F. Aloufi
Information Systems department
College of Computer Science and Engineering at Taibah University
Medina, Saudi Arabia
oofi@taibahu.edu.sa

Karim Djemame
School of Computing
University of Leeds
Leeds, United Kingdom
K.Djemame@leeds.ac.uk

Faisal Saeed
Information Systems department
College of Computer Science and Engineering at Taibah University
Medina, Saudi Arabia
fsaeed@taibahu.edu.sa

Fahad Ghaban
Information Systems department
College of Computer Science and Engineering at Taibah University
Medina, Saudi Arabia
fghaban@taibahu.edu.sa

Abstract— This paper presents the concept and characteristics of cloud computing, and it addresses how cloud computing delivers quality of service (QoS) to the end-user. Next, it discusses how to schedule one's workload in the infrastructure using technologies that have recently emerged such as Machine Learning (ML). That is followed by an overview of how ML can be used for resource management. Then, this paper aims to outline the benefits of using ML to schedule upcoming demands to achieve QoS and conserve energy. In addition, we reviewed the research related to ML methods for predicting workloads in cloud computing. It also provides information on the approaches to elasticity, while another section discusses the methods of prediction used in previous studies.

Keywords— Cloud Computing; Optimising Quality of Service (QoS); Resource management

I. INTRODUCTION

Cloud computing (CC) has been one of the most fundamental technology in the last few decades due to the changes which it made in the computing field. Historically, CC was emerged in early of 2006 when huge companies such as Amazon and Google started to provide people accessing to files via web instead of their desktops [1]. Accordingly, CC has different definitions from different scholars; for example, CC is defined in [2] as "Cloud computing creates a network-based environment vision to the users, which paves the way for the sharing of calculations and resources regardless of location". Another example of the definitions, the National Institute of Standards and Technology (NIST) [2] defined CC as a frame or model that is enabling universal to pool of computer resources such as storage, applications, services and shared data, which is available and released with less effort or interactions from the providers. In this sense, CC has been invented to meet the user requirements and to satisfy their needs in simple ways. Furthermore, one of the key characteristic of CC is elasticity, which is a feature in CC that is seeking to meet the needs of the user's with no interruption at run time.

It could be known that there are some traditional approaches to do elasticity, which is a key for QoS in the context of CC for optimizing the QoS to the end-user or the providers of CC. The approaches of elasticity have been explored and investigated with different modelling such as [3]. Even though mathematical modelling have done a forward step in meeting the user's needs, there would be a lack in the optimisation of QoS. This paper puts forward that there is a need to implement ML algorithms rather than math modelling to purse the QoS for CC.

II. CLOUD COMPUTING MODELS AND TRENDS

A. *Virtualisation*

Cloud resources, such as servers, network components and storage units, are components abstracted by virtualisation; therefore, CC relies heavily on virtualisation. In other words, the ability to run different operating system (OS) on one physical host (PH). In this context, the virtual machines (VM) will give the impression that they are accessing the hardware resources of PH, but that resources are actually being shared. According to Malhotra et al. [4], virtualisation makes off-site resources like applications and storage appear as though they were part of the device that a person is using. Thus, virtualisation is particularly interesting because it can consolidate several systems running in a VM. That is, virtualisation enables several VMs to run efficiently on one PH. As mentioned in [4], virtualization is primarily used to manage workload by making traditional computing more flexible, effective and economical. Virtualisation is applied at the level of hardware, such as VMs and containers (see section VI). A VM seems like a physical machine, and it provides a working environment that can run or host a guest's OS. Thus, the cost of running several VMs in one physical server is far less than that of using a server for each one. CC has exploited this technology to reduce the cost of the underlying hardware and to save energy. Figure. 1, which is taken from [5], illustrates the virtualisation layer on Cloud Computing.

978-1-6654-1224-7/21/$31.00 ©2021 IEEE

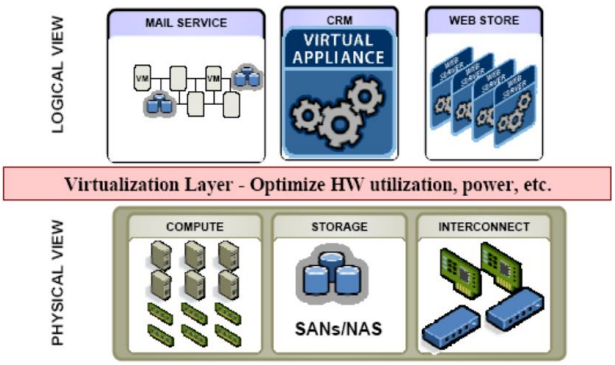

Figure 1. Virtualization [5]

B. *Cloud Computing*

CC is a relatively recent advanced technology which has changed the concept of computing considerably. CC is defined by the U.S. National Institute of Standards and Technology (NIST), as the pooling of resources to be shared appropriately to meet the requirements of users with respect to accessing resources, storing data and processing data [6]. Accordingly, it is possible that many organisations, academia or industries can take advantage of the concept of CC to facilitate their works and the services they provide to the end-user. In addition, QoS has become one of the key pursuits in the context of CC. CC can meet user's requirements and satisfy their needs while minimising the provider's costs for power consumption by applying such approaches to resource management.

Cloud computing provides services to consumers using on-demand and pay-as-you-go models [7]. NIST in [8] identifies five characteristics of cloud computing: on-demand services, broad network access, resource pooling, rapid elasticity and measured service. On-demand services refers to the ability of the cloud to support the end-user according to his demand for networks resources, such as storage and access to databases, without involving human contact. Broad network access allows users to utilize cloud resources over a network using a standard mechanism; the use of a heterogeneous component is promoted through these characteristics. Resource pooling is one of the primary features of cloud computing in which multiple consumers have access to multiple resources, such as storage, memory and processing. Resource pooling uses a multi-tenant model with diverse physical and virtual resources. Resources are dynamically assigned and reassigned by the cloud provider according to the end-users' requests. Rapid elasticity is also a core characteristic of workload prediction in cloud computing. It refers to the cloud's ability to be automatically provisioned to satisfy the needs of customers. Taking advantage of rapid elasticity is an efficient way to pursue high QoS and decrease power consumption. Elasticity allows CC to meet the user's needs with no interruption at run time [3], positively contributing to the QoS. Rapid elasticity will be considered to explore ML to predict the workloads of servers. Measured service involves calculating the resources used by consumers based on consumer utilisation. Measured service is an important feature of the cloud because it allows the needs of the consumer to be satisfied while meeting the energy conservation requirements of the provider.

III. Cloud Computing Service Models

Cloud computing systems have three models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). These are known as the taxonomy of cloud systems. Accordingly, the critical point about cloud computing as a technology is that everything is seen as a service; that means hardware is a service, function is a service and the database is a service. The models are explained further in the following sections.

A. *Software as a Service*

Software as a Service (SaaS) utilises the cloud to give users access to the software. This means that the software runs on the cloud, not on the customer's machine. SaaS brings those applications running in the infrastructure of the cloud to the end-user. Such applications can be accessed either directly by the end user or through a program interface on a network, thus removing the burden from the underlying cloud infrastructure. Apps like Google Docs are an example of SaaS.

B. *Platform as a Service*

Platform as a Service (PaaS) provides computing capabilities to developers for building applications and then deploying them on the cloud infrastructure. PaaS allows developers to design and control their applications, but it does not allow them power over the underlying infrastructure. This means that the user (developer) does not require the software to be stored on the cloud; rather, the user uses the platform to deploy the software. The cloud, therefore, provides the user with a software stack—a platform—where the user can run and deploy the software, such as Microsoft Azure. Thus, the difference between SaaS and PaaS is that the former hosts completed software and the latter hosts the environment that allows for the development of software.

C. *Infrastructure as a Service*

Infrastructure as a Service (IaaS) provides the infrastructure of a real machine to the customer, allowing them to directly use cloud resources, such as network and storage components, and the virtualisation technology makes that happened seamlessly. Essentially, IaaS provides a customer with access to the server, usually virtually, after which the customer owns that machine and can deploy, run and manage apps. It is important to note that although IaaS provides the resources in which the customer can deploy their apps and software, the customer does not have power over the underlying resources since they are being abstracted by virtualisation. However, IaaS ensures that customers have full control over the operating systems and the deployed applications. Amazon elastic compute cloud (EC2) is an example of an IaaS.

IV. Cloud Computing Deployment Models

In cloud computing, there are three models of deployment, each with their own characteristics and definition based on the infrastructure, their features and the level of control. The three models are private, public and hybrid. Choosing the appropriate deployment model depends on the requirements of the user, such as the cost, the needs, the level of privacy and the kind of service being deployed.

A. *Private Cloud Model*

The private cloud model is provided for specific purposes to a single entity (organisation) consisting of several consumers. It can be managed internally or by a third party

depending on whether it is on-premises or off-premises [8, 10]. As the name implies, a private cloud infrastructure means that services are accessed and managed by a single entity. The benefits of this model make it appropriate for government agencies and organisations using sensitive data. Such groups often prefer private cloud infrastructure because of the small amount of data and the need for it to be stored securely. Compared to the other deployment models, the private model is more expensive and requires much effort to be deployed.

B. Public Cloud Model

The public cloud is the model most used by cloud consumers. This is the dominant cloud deployment model because public consumers utilise it, and this could clarify that the provider of services will own the public cloud with their own policies [10]. The public cloud is accessed over the Internet through either a pay-as-you-go subscription or a contract. The advantage of this model is that it allows organisations to focus on running their business instead of the infrastructure. Also, public consumers can take advantage of the pay-as-you-go feature to request more (or less) resources as needed (scale up and scale down). Accordingly, the public cloud may play a pivotal role in minimising the cost of infrastructure and operation for organisations. Public cloud providers must commit to the requirements in the Service Level Agreement (SLA) between them and the consumer. The only disadvantage of the public cloud is the lower level of privacy and security it provides, compared to other models.

C. Hybrid Cloud Model

The hybrid cloud model is a mix of public and private clouds. This model provides organisations with the flexibility to have more resources, especially in handling the peak load. The hybrid cloud model can play an essential role in preventing cloud bursting. This occurs when the number of entities reaches the maximum capacity for service. At such times, entities can be offloaded to the public part of the hybrid model. Entities utilise this model for its benefits in both cost and security.

V. Resource Management and Scheduling

Resource management is considered one of the major challenges in CC because of the complexity and heterogeneity of the system. According to Kumar and Manoj [11], resource management is the method of distributing demands, such as storage resources, for cloud providers and cloud consumers. Cloud resource management is also affected by massive interactions, some of which are unpredictable, such as failure of the system. Another challenge for cloud providers involves elasticity, particularly that from a fluctuating large load. The decisions on resource utilisation must be made using accurate measurements of the physical and virtual resources needed to distribute applications [11]. A cloud service provider tries to fulfil the requirements of customers, but this requires complex policies and decisions; resource management requires optimisation of multiple objectives, such as load balancing, energy usage, costs, utilisation of processors and availability of machines.

Scheduling is about deciding how to allocate system resources, such as Central Processing Unit (CPU), memory, disk, storage, and network bandwidth, etc. It could be said that scheduling has become one of the most significant issues for CC. The researchers in [12] stated that 'scheduling algorithms should order the jobs in a way where balance between improving the performance and quality of service and at the same time maintaining the efficiency and fairness among the jobs'. Hence, resource scheduling is about efficiently assigning jobs to be run on machines.

Resource management and scheduling and their roles in CC have been outlined. Thus, resource management and scheduling, particularly when used with ML, play a significant role in optimising QoS and reducing power consumption.

VI. Trends in Cloud Computing

The use of the cloud computing paradigm has rapidly increased, and new trends have emerged in the cloud industry. According to [13], as cloud computing matured, advancements appeared in the underlying technologies, such as containers. Other developments arose from these advancements, such as edge computing, serverless computing, and evolving trends in Information and Communication Technology (ICT), such as ML. These key technologies are presented in the following sections.

A. Containers

Container technologies became of interest to Internet companies with the emergence of Docker [13,14]. Containers are lightweight in comparison to a VM since a VM brings the guest's OS with it, whereas containers rely on the core OS of the host machine. It is also possible to associate containers with microservices due to the ability of containers to isolate specific codes to be executed. Because containers are lightweight, they offer quick start-up and require little memory. Therefore, they require a small number of resources [13], enhancing the performance of applications in comparison to VMs. There is still a need for VMs to execute containers because containers must be run on the OS of VMs, and containers can be implemented in cases such as batch computing and microservices where the best performance is necessary.

B. Edge computing

Edge computing (EC) brings the cloud closer to the end of the network device. In other words, EC performs tasks on behalf of cloud services at the edge of the network. Therefore, in real-time analysis, latency becomes minimum and the bandwidth becomes larger and more available. in this sense, if a massive amount of data are generated at the edge of the network, it is more effective to process those data there instead of sending them to the cloud [13]. EC also helps to solve issues of end-device mobility. For example, autonomous vehicle and technologies used in the Internet of Things (IoT) use EC to process the massive amounts of data they produce. This would help to avoid bandwidth, network and latency issues. EC is also useful for CC since it provides the opportunity to offload part of the workload from the cloud to the edge [13]. Thus, EC plays a vital role in reducing the response time in some scenarios, such as autonomous vehicles, and in reducing the burden of workload on cloud resources.

C. Serverless computing

Serverless computing means that the software architecture hides the server from developers. Serverless computing reduces the need to deal with the backend code and it plays a pivotal role in resource management in the cloud [13]. Serverless computing has contributed to eliminating the need for immersing hardware and software and to reducing the total cost paid to the resource. This means that serverless computing is a technology trend that offers new opportunities to offload part of the application's logic far away from the core

system [13], providing simplicity, speed and flexibility for developers.

D. Machine Learning

CC can embrace a vast amount of data because of the existence of power computing. Thus, ML assists experts and developers in their jobs in the cloud instead of their local machines. ML is useful to improve the mechanism of resource allocation, optimise the usage of resources and minimise the use of energy. The benefits of ML have attracted researchers and participants to apply it to the cloud, such as in [13], where it was concluded that the optimisation of resources could be increased by implementing ML techniques. Clearly, applying ML to resource allocation and task scheduling contributes positively to meeting SLAs and reducing power consumption.

VII. Service Level Agreements (SLAs)

A SLA is a contract between the provider and customer regarding the provision of service. The properties of an SLA are either functional or non-functional. Functional properties refer to what the client must have to access the IaaS (e.g. a VM) and the number of cores required. Non-functional properties refer to things such as security. For example, the SLA may allow the client to request a VM that runs on a server that is not shared with other VMs. Hence, QoS concerns the system's capability to fulfil the service requirements. As stated in [15], QoS is a part of the SLA required to be enforced. The nature of the cloud allows for monitoring the QoS as it is written in the SLA. SLAs contain information on the availability of resources, the reliability of service components and other warranties for each party. In some cases, there is a penalty for the service providers if, for instance, the they breach the SLA in providing cloud services. Thus, SLAs protect the rights of both parties.

VIII. Predicting workloads and optimising QoS in CC

Several studies have contributed to the information on predicting workloads in the cloud, optimising QoS for cloud consumers and reducing power consumption in cloud computing. This section presents the previous work applying models to predict workloads and the data being used and their achievements in this domain. Then we consider the most important studies published from 2016 to 2020 that looked at predictions using ML techniques or mathematical modelling. We take the study by Islam et al. [16] as an example of prediction methods for resource management and strategy provision.

A. Elasticity in Clouds

Efficiently exploiting the elasticity of a cloud is critical for the instantaneous provision and de-provision of resources. Elasticity is essential to achieve high performance in the cloud. Workloads must be predicted to avoid scaling delay and to improve QoS [3]. Previous studies have used two methods for auto-scaling resources: proactive and reactive. The proactive method applies ML techniques to predict upcoming workloads to efficiently provision resources. Hence, in order to improve the elasticity mechanisms in the cloud, ML algorithms must be applied to predict workloads more accurately and, subsequently, to assign them with the scheduling algorithm. This allows for excellent and efficient provisioning or de-provisioning of cloud resources, improving QoS and minimise power consumption. CC, by its nature, changes state frequently; therefore, exploring ML algorithms to exploit elasticity supports the drive to satisfy customers and minimise the total cost of power consumption for cloud providers.

B. Predicting workloads

Predicting workloads is fundamental to provisioning resources efficiently. Provisioning resources in the cloud to accomplish different objectives, such as improving QoS and minimising power consumption, has been broadly studied. Researchers have tried to predict workloads using different approaches; for example, the authors in [3] applied mathematical models to predict demands to solve delays in scaling.

Islam et al. [16], developed a prediction-based model for resource management and strategy provision using neural networks and linear regression. The aim of the model was to solve the issue of delays in allocation by anticipating clients demand. They approached ML techniques with respect to time, applying two algorithms: error correction neural network (ECNN) and linear regression. They justified using these algorithms because they are effective for forecasting [16]. The algorithms in that study were used on the dataset of CPU usage collected by the TPC-W benchmark. They generated and emulated numbers from sessions by users in online shops. Data were collected and the CPU utilisation was considered for prediction by the proposed model. The sample CPU utilisation as a dataset was used to train the proposed model in [16] to predict the usage of a resource correctly. The researchers completed the experiments with and without a sliding window. The model showed promising results that revealed greater success from using the neural network model with a sliding window for estimating resource usage in the cloud.

The proposed prediction strategies provided efficient ways to adapt the resources in the cloud in terms of performance and cost. In the same way, it is important strives to reduce power consumption and achieve great QoS. This would also improve the provisioning of resources.

Wang et al. [3], proposed a new trigger strategy for provisioning resources to improve QoS and meet the user's needs as they appeared in an SLA that involved an automatic-scaling mechanism. That study took three approaches: a time series approach using three models (MA, AR and ARIMA); 2) the Kalman filter and 3) a pattern-matching model. The proposed trigger strategy in [3] was based on a pattern-matching model, while the other triggers depended on the threshold approach. Wang et al. [3] predicted the workload by monitoring the data of CPU utilisation using Aliyun VMs as tools. To get the first dataset (a CPU workload time series), they ran a regular word count program. They then gathered the workloads of the computational tasks, web applications and applications for memory consumption. Afterwards, they transformed the workloads into strings of historical patterns (scaling-up strings, scaling-down strings and stable strings). They used the mean absolute percentage error (MAPE) metric to evaluate the results, which showed improved prediction accuracy and a reduction of delays in the automatic scaling. However, even though the mathematical modelling in that study was an advanced step in provisioning resources, there was still a need to implement ML techniques to predict the workload and achieve QoS, and there was a need to improve the accuracy of the prediction. For those reasons, we can use a standard dataset from Google Cloud Trace.

While we have goals like those of Wang et al. [3] which is to improve QoS and reduce the power consumption, we can use ML techniques to predict workload.

Bin et al. [9] tried to predict the precise level of demand for cloud resources. They considered the planning of cloud capacity as a classification problem. They also proposed an integrated framework that forecasts changing demands to minimise the cost of providing cloud resources. They used piecewise linear representation (PLR) to classify a time series of cloud resource requirements to identify the changing trend for each duration, then they used weighted SVM to match each period's statistical information with its label and to forecast the trend of the next period. Finally, they used an incremental learning method to make sure that the model updated at a low cost using incoming requests. They used the IBM smart cloud enterprise (SCE) trace dataset, which included 48,368 records created by 2,024 users over five months in 2011. They evaluated their results by applying PLR to two other traditional time series segmentations—sliding window and bottom-up—and by comparing weighted SVM with several classifiers, such as the k-neighbours classifier.

They showed that their model could customise the degree of changing demands and the importance of different types of changing trends to reduce the overall cost of provisioning. The segmentation strategy of a time series still has some limitations, such as threshold selection and the unknown relationship between the threshold and the degree of changing cloud demands. It may also be possible to integrate regression and classification approaches to increase the accuracy of the predictions.

It is needed to reduce power consumption by the data centre. That would reduce the costs of over-provisioning as well. Therefore, SVM regression techniques can be used to predict upcoming demands. Moreover, the dataset from Google Cloud Trace, which is very large, can be used similar to the huge number of records used by Bin et al. [9].

Kumar et al. [17] developed a model based on ML techniques and neural networks. They combined a neural network with a self-adaptive differential evaluation to predict demand, to improve QoS and to avoid any violations of the SLA. They used a prediction model that extracted the requests and mapped them onto a time unit level. The neurons had to be trained to produce better predictions. This model was even considered to be an effective in its domain [17]. Those researchers used data from Saskatchewan and NASA HTTP traces, and the model was trained by a neural network with self-adaptive differential evolution (SaDE). It obtained better results with 10 inputted neurons. Their results showed enhanced accuracy. The researchers determined that their model reduced both the number of violations of the SLA and operational costs. The time interval used in that study was one minute. Their outcomes indicated that the proposed approach should further explore ML techniques such as SVM.

In addition, Kumar et al. [17] improved QoS by avoiding SLA violations. And Montero et al. [18] looked at the problems of achieving QoS, such as long response time, particularly during high traffic loads and fluctuating demands. They found that the SVM model they used to predict the upcoming demand provided an optimal and unique solution. This is because SVM is a global solution, whereas artificial neural networks (ANNs) might suffer from local minima. The researchers proposed a novel mechanism to guarantee QoS that provided an optimal number of resources during peak-demand and decreased resource over-provisioning to save power and decrease the total cost of the infrastructure. This study used a time series approach and forecasted using ML techniques (SVM). The proposed mechanism in [18] was based on a proactive (predictive) time series mechanism. Montero et al. [18] forecasted workload based on historical observations of a web server. Their proposed method estimated the optimal resources needed to ensure QoS and to reduce over-provisioning. They implemented an SVM technique using different functions of kernel, such as a normalised polynomial kernel and a polynomial kernel, and they also applied different configuration constraints to obtain optimal results. Their data were collected over a four-week period of exactly 672 hours. They chose the lag variable of 24 hours and did the experiments in the last hour. Their results showed a close-to-optimal allocation of resources. However, the study showed a clear need to apply ML (SVM) techniques to predict the workloads of big data clusters such as Hadoop or Spark. Our objectives are like those of Montero et al. [18]: to improve QoS and reduce the over-provisioning of resources causing power consumption.

In addition, there are many studies in authors in dynamic resource allocation and scheduling in cloud computing. For instance, the authors in [19] proposed a solution for dynamic resource allocation that provides an energy efficient VM architecture for cloud computing. In [20], the authors investigated and empirically compared some of the most recent scheduling heuristics in cloud computing.

Finally, in this paper, we recommend to use SVM techniques to predict workloads in the upcoming demands. At the end of this section, the above literature review is summarised in the Table 1.

Criteria / Paper	Islam et al. [16]	Wang et al. [3]	Bin et al. [9]	Kumar et al. [17]	Montero et al. [18]
Data used	Dataset of CPU usage	CPU workload time series	IBM Smart Cloud Enterprise (SCE) trace data dataset	Dataset of Saskatchewan and NASA HTTP traces	Real web service logs from the Complutense University of Madrid
Prediction /ML techniques/ math model	Neural network and linear regression	Math model	WSVM	ML techniques and neural network	SVM
Results related to QoS	Promising results in estimating resource usage in the clouds	Contributed by reducing scaling delay	Not considered	Reduced violations in the SLA	Optimised QoS
Power consumption	Not considered	Not considered	Not considered	Reduced operational costs in data centres	Reduced over-provisioning
Limitations/ Further works	Implement and evaluate their models for different workload generators	Need to apply ML techniques	Several ideas such as exploring the relationship between threshold and the degree of cloud changing demands	Explore further techniques such as SVM with the same level of unit	Predicted workloads of big data cluster such as Hadoop or Spark clusters
Year	2012	2016	2018	2018	2019

IX. Conclusion

This paper has summarized the critical concepts of cloud computing and identified the future trends of this field, including containers, edge computing, serverless computing and machine learning techniques as they pertain to cloud computing. Some concepts, such as resource management, scheduling and SLAs, were defined to provide an understanding of the issue of QoS and the allocation of workloads. In addition, this paper has covered recent studies that implement ML techniques to conduct the predictions of upcoming workloads at the level of the cloud data centres and their contributions to mitigate the issues of QoS and power consumption.

The literature review has started by outlining the features of elasticity and how it is vital to discover ML techniques to be implemented in resource management at the level of the cloud data centre. A review of the literature highlighted the methods used by other researchers to predict the demands at the level of the cloud to manage resources efficiently, and it indicated the importance of ML techniques that are still in the early stages of exploration. Even though the researchers have studied ML techniques to predict workload, it was convenient for further investigation to explore ML algorithms and implement those algorithms in a real cloud data centre. This paper has concluded that the results of previous efforts were promising for finding ML techniques and implementing them in real clouds data centres.

In conclusion, Islam et al. [16] had promising results when they used ML algorithms to predict the workload. Their results showed how many resources would be used in clouds and they were a leading key for further studies. They noted that it could be possible to accommodate other ML methods to predict the workload for CPU utilisation by SVM.

Therefore, this paper discussed how possible to optimise QoS and predict the workload; therefore, a novel algorithms and real data like Google datasets for CPU utilisation can be used to conduct real experiments with real data in real environment.

References

[1] Antonio Regalado, (2019). Who coined 'Cloud Computing'?. [online]. [Accessed 18 February 2021]. Available at: https://www.technologyreview.com/2011/10/31/257406/who-coined-cloud-computing/

[2] Nalini Subramanian, Andrews Jeyaraj, "Recent security challenges in cloud computing", *Computers & Electrical Engineering*, Print ISSN 0045-7906, pp. 28-42 , Vol. 71, DOI: 10.1016 /j.compeleceng.2018.06.006, Available: https://www.sciencedirect.com/science/article/pii/S0045790617320724

[3] Yazhou Hu, Bo Deng, Fuyang Peng and Dongxia Wang, "Workload Prediction for Cloud Computing Elasticity Eechanism", in *Proceedings of the IEEE International Conference on Cloud Computing and Big Data Analysis 2016 (ICCCBDA)*, 2016, Chengdu, China, DOI: 10.1109/ICCCBDA.2016.7529565, pp. 244-249.

[4] Malhotra, Lakshay, Devyani Agarwal, and Arunima Jaiswal, "Virtualization in Cloud Computing", *Information Technology & Software Engineering*, 2014, 4(2), pp.1-3.

[5] Djemame, K. "Virtualisation". COMP580 Cloud Computing. University of Leeds, 2020.

[6] Badger, Mark Lee, Timothy Grance, Robert Patt-Corner, and Jeffery M. Voas, "Cloud Computing Synopsis and Recommendations", *National Institute of Standards & Technology*, 2012, pp. 1-81.

[7] Lima, H., Aragão, F., Lima, J., Sousa, F.R. and Monteiro, J.M. CloudSimDB: Um Simulador para o Provisionamento de Máquinas Virtuais para o Processamento de Aplicações Centradas em Banco de Dados. In: 28th Brazilian Symposium on Databases, p.7.

[8] Mell, P. and Grance, T. The NIST definition of cloud computing. Special Publication 800-145. Gaithersburg: National Institute of Standards and Technology. 2011, pp.1-7.

[9] B. Xia, T. Li, Q. Zhou, Q. Li and H. Zhang, “An Effective Classification-based Framework for Predicting Cloud Capacity Demand in Cloud Services”, in *IEEE Transactions on Services Computing,* 2018, DOI: 10.1109/TSC.2018.2804916, pp.1-13.

[10] T. Dillon, C. Wu and E. Chang, “Cloud Computing: Issues and Challenges”, in *Proceedings of the IEEE international conference on advanced information networking and applications* 2010, 20-23 April 2010, Perth, WA, Australia, DOI: 10.1109/AINA.2010.187, pp.27-33, Published by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/5474674

[11] Kumar, ER.Manoj, “Cloud Computing in Resource Management”, *International Journal of Engineering and Management Research (IJEMR*). 2018, 8(6), pp.93-98.

[12] Kaur, D. and Sharma, T. Scheduling Algorithms in Cloud Computing. *International Journal of Computer Applications.* 2019, 975, pp.16-21.

[13] Buyya, Rajkumar and Srirama, Satish Narayana and Casale, Giuliano and Calheiros, Rodrigo and Simmhan *et al.*, “A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade”, in Association for Computing Machinery, 2018, Vol. 51, DOI: 10.1145/3241737, pp.1-51.

[14] Merkel, D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal.* 2014, 2014(239), p.2.

[15] Patel, Pankesh, Ajith H. Ranabahu, and Amit P. Sheth, "Service level agreement in cloud computing", 2009, Available: https://corescholar.libraries.wright.edu/knoesis/78/

[16] Sadeka Islam, Jacky Keung, Kevin Lee, Anna Liu, “Empirical prediction models for adaptive resource provisioning in the cloud”, *Future Generation Computer Systems.* 2012, Vol. 28, pp.155-162. Available: https://www.sciencedirect.com/science/article/pii/S0167739X11001129

[17] Jitendra Kumar, Ashutosh Kumar Singh, “Workload prediction in cloud using artificial neural network and adaptive differential evolution”, *Future Generation Computer Systems.* 2018, Vol. 81, pp.41-52. Available: https://www.sciencedirect.com/science/article/pii/S0167739X17300444

[18] Moreno-Vozmediano, Rafael Montero, Rubén S.Huedo, Eduardo Llorente, Ignacio M., “Efficient resource provisioning for elastic Cloud services based on machine learning techniques”, *Journal of Cloud Computing*. 2019, Vol. 8,DOI: 10.1186/s13677-019-0128-9, p.5.

[19] D. Alsadie , Z. Tari, E. J. Alzahrani, and A. Y Zomaya. Dynamic resource allocation for an energy efficient vm architecture for cloud computing. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1-8). 2018.

[20] M. Ibrahim, S. Nabi, A. Baz, H. Alhakami, M. S. Raza, , A. Hussain, and K. Djemame. An in-depth empirical investigation of state-of-the-art scheduling approaches for Cloud computing. IEEE Access, 8, 128282-128294. 2020.